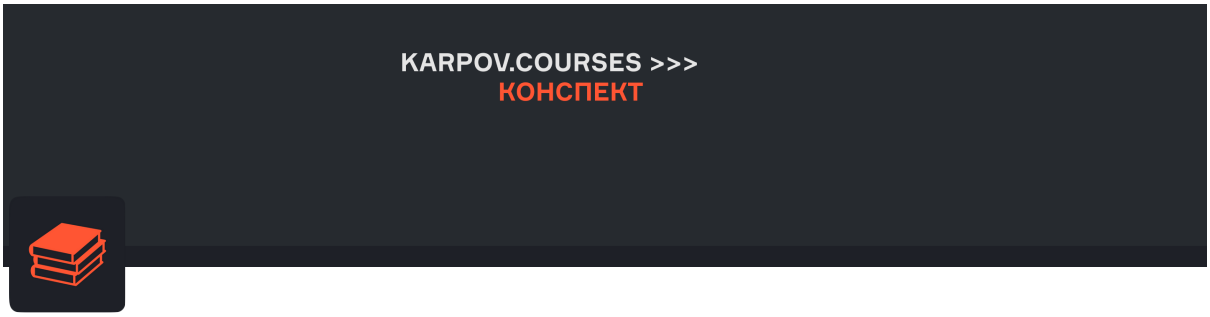


> Конспект > 2 урок > Метрики в A/B-тестировании

> Оглавление

> Зачем нужны метрики

> Проблема выбора

Ситуация: хотим улучшить дизайн сайта онлайн-магазина, предполагая, что клиентам станет удобнее им пользоваться. Ожидаем, что это приведёт к увеличению продаж.

Можем провести пилот, разделив пользователей на две группы, а после рассчитать метрики интернет-магазина:

- Количество кликов по разным кнопкам и баннерам;
- Количество просмотров страниц;
- Количество добавлений товаров в корзину;
- Количество покупок;
- Средняя стоимость покупок;
- Количество товаров в заказе и т.д.

Одни увеличились, вторые уменьшились, третьи остались без значимых изменений. Например, заказов стало больше, средний чек упал. Что делать? Как понять, был ли успешен пилот?

Если заранее не зафиксировать метрику, то постфактум результатами можно манипулировать в некотором роде. Нужно определить, как принимать решение на основе всего множества метрик.

> Конкретная гипотеза

Протестировать можно только чётко сформулированное предположение:

- Что изменим;
- На что это повлияет;
- Как измерить.

Гипотеза: если изменим X, это повлияет на Y, можно измерить метрикой Z.

Примеры:

1. Теряем клиентов из-за больших очередей (актуальность проблемы);
2. Своевременно выводить кассиров в час пик (изменение);
3. Очереди станут меньше, покупок больше (влияние);
4. Продажи выросли (способ измерения).

> Целевая метрика

Целевая метрика должна хорошо отражать конечную цель изменений (в идеале — совпадать с основной метрикой бизнеса, обычно это деньги).

Использовать общую выручку как метрику не всегда возможно, т.к. изменения могут быть незаметными при работе с небольшой группой товаров (например, фрукты и овощи). Хотя изменение выручки только по группе заметить можно.

Иногда стоит взять более специфичную метрику, но не стоит углубляться слишком сильно, т.к. процесс может повлиять на что-то ещё. Для локальных изменений специальные метрики более чувствительные.

Пример: если меняем стеллажи для фруктов и ожидаем, что они станут продаваться лучше, то можно смотреть на продажи именно фруктов.

Каннибализация — сокращение продаж одного товара вследствие увеличения продаж другого товара.

Пример: после замены стеллажей продажи фруктов выросли, а овощей упали.

Если бы две категории из примеров (овощи и фрукты) рассматривались совместно, можно было избежать каннибализации.

Лучше не использовать метрики, которые являются прямым следствием изменений.

Пример: добавили кнопку и по ней начали кликать.

Очевидно, что такая метрика должна увеличиваться, но она не влияет на целевые показатели кампании.

Вернёмся к примеру с очередями: возможно, стоит использовать среднюю длину очередей как целевую метрику кампании по привлечению клиентов. Даже если выручка не увеличилась, но длина очередей упала, это может стать инструментом для долгосрочного развития.

> Иерархия метрик

Целевую метрику можно разбивать на составляющие по различным правилам (например, по возрастным когортам), эффект в которых будет виден лучше. Иерархия метрик обозначает связи между метриками и раскладывает ключевую метрику на составляющие. Она позволяет более точно установить связи между ними и точнее отслеживать изменения.

Пример иерархии метрик для выручки

> Свойства метрик

То, насколько хорошо метрика отражает цель:

- Хорошо (LTV);
- Средне (средний чек) — клиент может делать большие покупки, но редко;
- Плохо (CTR, *click-through rate*, клики на баннер) — люди могут кликать на рекламу, но не совершать покупки.

Скорость ответа:

- Медленно (LTV) — чтобы честно посчитать LTV, необходимо дождаться, пока клиент перестанет пользоваться нашими услугами. Например, услугами онлайн магазина клиенты могут пользоваться годами, поэтому чтобы измерить LTV, нужно долго ждать;
- Средне (покупка) — можно смотреть на совершение целевого действия во временном окне: например, совершение покупки в течение недели после просмотра рекламы;
- Быстро (CTR) — получаем ответ почти мгновенно.

Часто скорость и то, насколько хорошо метрика отражает цель, имеют обратную зависимость.

Достоверность:

- Высокая (покупка) — подтверждается транзакциями; если человек купил товар, то, видимо, товар ему понравился;
- Средняя (просмотр) — задержка на материале более N секунд считается заинтересованностью, хотя на деле пользователь просто отвлекся;
- Низкая (отзывы) — люди, у которых всё хорошо, редко пишут отзывы.

Множество возможных значений:

- Бинарная (переход по ссылке) — из-за своей простоты имеет определённые плюсы: легкость вычисления, более высокая мощность тестов и т.д.;
- Дискретная (количество товаров в корзине);
- Непрерывная (длина сессии).

Необходимо заранее фиксировать политику работы с выбросами в данных.

Богатство значений:

- Богатая (длина сессии) — каждый пользователь сайта генерирует какое-то значение;
- Средняя (CTR);
- Скудная (покупка редких товаров) — для получения статистически значимых результатов потребуется много данных.

Способ вычисления:

- Среднее;
- Отношение;
- Квантиль.

> Синтетические метрики

На основе имеющейся метрики при помощи некоторых преобразований можно сделать другую метрику, которая будет обладать другими свойствами. Такие метрики называются синтетическими.

Пример: бинаризация небинарной метрики — из непрерывной можно сделать бинарную.

Мы хотим увеличить количество товаров в заказе, имеющаяся метрика — среднее количество товаров в заказе. Предположим, в среднем сейчас по одному товару. Можем заменить среднее количество товаров долей заказов, где больше одного товара (не считать количество товаров в конкретном заказе, а отмечать его флагом, если он соответствует заданному условию по количеству товаров).

Тогда среднее количество товаров в заказе -> доля заказов, где больше одного товара.

Мы теряем часть информации: мы не будем знать, на сколько больше одного товара было в заказах, зато теперь не нужно обрабатывать выбросы и можно использовать тесты для бинарных метрик, что может увеличить чувствительность тестов.

Ещё один пример с долгой и точной метрикой: мы хотим измерять LTV, но нужно ждать пока пользователь уйдёт от нас. По историческим данным можно построить модель, прогнозирующую LTV пользователя, и использовать её прогноз в качестве метрики.

> Вспомогательные и контрольные метрики

> Вспомогательные

Вспомогательные метрики помогают понять, что пилот идёт так, как задумывался.

Примеры вспомогательных метрик для "наблюдения" за страницей с новой акцией на сайте:

- Просмотры — если ссылка в сети неверная, то можно на раннем этапе выяснить проблему;
- Клики;
- Добавления в корзину.

Обычно целевой метрикой при проведении акции являются суммарные продажи, т.к. товары с промо могут каннибализировать другие позиции.

Если вспомогательные метрики не соотносятся с целевой и говорят, что продажи промо товаров идут плохо, то стоит задуматься, что же является

драйвером продаж. Возможно, рост продаж никак не связан с промоакцией.

В качестве вспомогательных метрик лучше использовать быстрые, не слишком скудные метрики, чтобы получать "опережающие" целевую метрику сигналы.

> Контрольные

Контрольные метрики позволяют отлавливать проблемы. Пополняются по мере выявления новых проблем после экспериментов.

Примеры:

- Обращение в саппорт;
- Возвраты;
- Время загрузки страницы.

> Случайные эффекты

Может случиться так, что по целевой метрике эффекта нет, а по вспомогательной или контрольной — есть.

Пример: продажи не выросли, но значимо уменьшилось количество обращений в саппорт — является ли это следствием изменения или это случайность?

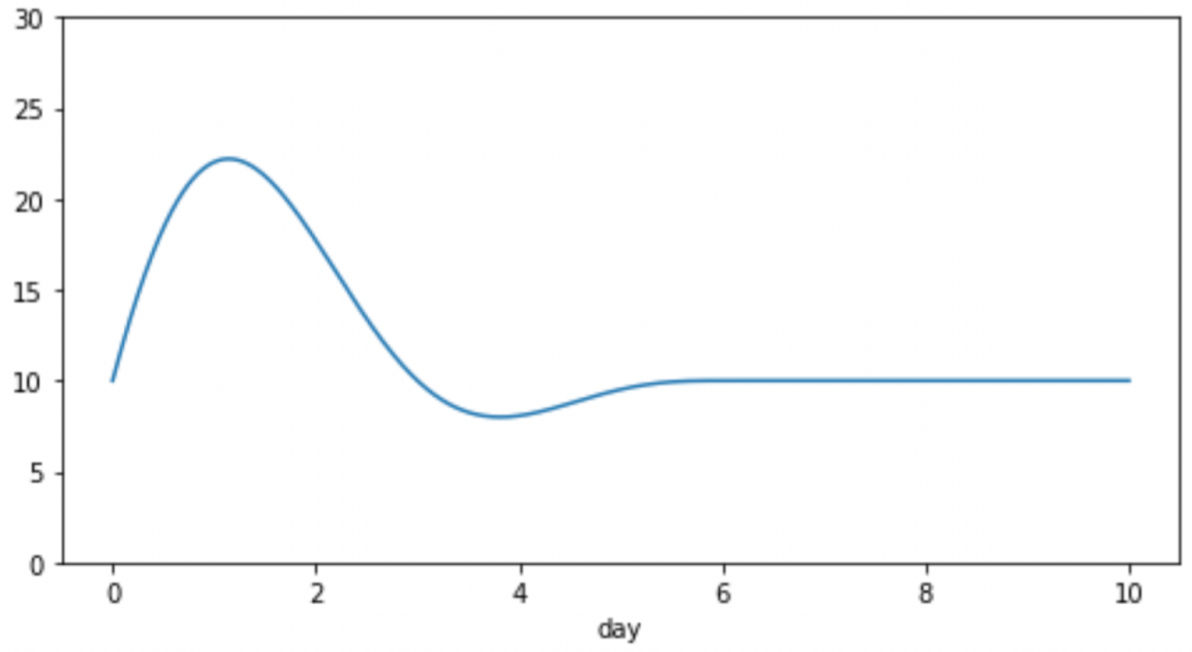
При уровне значимости 90 процентов, 10 метрик из 100 покажут значимое улучшение, даже если ничего не менять.

- Для подтверждения эффекта по нецелевой метрике лучше провести отдельный эксперимент.
- Можно делать поправку на множественную проверку гипотез.
- Подробнее можно узнать из лекции про множественное тестирование гипотез.

> Эффект новизны

После начала эксперимента может быть всплеск активности из-за любопытства, которое со временем пройдёт.

Чтобы нивелировать это явление, можно не учитывать первые дни после внесения изменений.



Пример эффекта новизны для некоторой метрики

> Процедура принятия решений

Решения всегда принимаются по заранее оговоренному алгоритму и в момент времени, определённый ещё до начала тестов.

Есть соблазн интерпретировать данные в свою пользу:

- Выбрать другую метрику;
- Установить другой уровень значимости;
- Удалить выбросы.

Гипотеза принимается, если:

- Ключевая метрика значимо улучшилась;
- Дополнительные метрики согласуются с гипотезой;
- Контрольные метрики не указывают на проблемы.

Полезно чётко фиксировать мотивацию принятия решения: что решили, почему решили так, дополнительные наблюдения и инсайты.

> Что делать после завершения теста

Тест положительный:

- Запланировать итерации с этой же гипотезой (поменять выкладку не только фруктов, но и других товаров);
- Перенести выводы на другие этапы воронки;
- Завершить линию экспериментов.

Тест отрицательный:

- Попробовать другое исполнение гипотезы;
- Внести изменение в гипотезу;
- Завершить линию экспериментов.
